Fast Solution

Protein Functional Sites Prediction (PFSP)

User Guide

Version 1.2 Multi Core

Thread safety powered by Intel® thread checker

Developer & Author: Dr. Ruan Xiaoyang

Correspondence to: ruansun@163.com

ruansun@yahoo.cn

IMPORTANT NOTE:

We strongly suggest you to carefully read this user guide before using PFSP.

PFSP DOES NOT support nucleotide sequence.

To reduce time to acceptable level in very time-consuming steps (such as estimation of pair wise Bayesian exact genetic distance for thousands of protein sequences), PFSP automatically FILTERS OUT ambiguous amino acid (AA) abbreviations such as *X* (for unknown amino acid residue*)*, *B* (for aspartate or asparagines), *Z* (glutamate or glutamine), *O* (pyrrolysine), **-** (gap of indeterminate length), * (translation stop).

DO NOT separate the PFSP main program from other shipped files in the original package on first time use.

Try to run PFSP on multi-core CPU system to gain a boost in speed.

Version Information

Version 1.0

Version 1.1

New iteration function was added to Multiple Sequence Alignment (MSA) module. The iteration function is composed of midway iteration and final alignment iteration, which significantly improves alignment quality by minimizing unreasonable matching errors without sacrificing much time.

A preview mode is added to enable users to preview the MSA result and make timely adjustment on parameters. Users can also directly refine a previously generated MSA result from main menu.

Pair wise evolutionary distance (and also variance) record was automatically saved. PFSP automatically reminds users of existed distance file when same number of sequence and distance estimation algorithm were encountered again. This is especially useful for repeated massive sequences alignment.

Results were now saved to subdirectories under C:\PFSP.

A few bugs were corrected.

Version 1.2 Multi Core

Some tedious algorithms were optimized to significantly boost the speed of MSA. Added support to Multi core CPU, so the performance can be further improved by running on Multi-cored system (2× faster on Duo core CPU system, 4× faster on Four core CPU system…).

Added support for MSA result generated by ClustalW program.

The bug of unable to run properly on non C:\ disks was corrected.

First Time Use

On first time use, the following files in PFSP package will be copied to C:\WINDOWS directory for next time use. PFSP will ask you to provide paths for these files if PFSP failed to find them in the same folder.

Blocks.dat [The block file used to generate BLOSUM scoring matrices (if you accidentally lost it, search http://blocks.fhcrc.org/blocks/uploads/blosum/ for it)]

_ProFunSit_BLOSUM_RltvFreq [Pre-calculated BLOSUM relative frequency file with re-clustering standard from 100% to 14%.]

_ProFunSit_t_s_thld [Pre-calculated BLAST hit score and extension score threshold]

PFSP_1.2_Userguide.pdf [User guide]

After first time use, you can copy/cut PFSP alone to other directory without above files

Log File

For each run of PFSP, a log file was saved to C:\ProFunSite.log, which will be automatically cleared the next time PFSP was initiated.

Table of contents

File Format

PFSP supports two kinds of format--**FASTA format** and **Custom format**.

FASTA format. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. "|" was used follow ">" to separate gi-number, sequence source, accession number and locus information. PFSP does not have limitation on number of characters in each line. A newline character is necessary to separate description line from sequence data. The sequence ends if another ">"appears. An example sequence in FASTA format is:

>gi|532319|pir|TVFV2E|TVFV2E envelope protein

ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRTQIWQKHRTSNDSRGTNDPK-RIFFQRQWG**B**DPETANLWFNCHGEFFYCKMDWFLN**Z**YLNNLTV
DA**O**DHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKKTYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGFAPTEVRRYTGGHERQKRVPFV
XXXXXXXXXXVQSQHLLAGILQQQKNLLAAVEAQQQMLKLTIWGVKNLLAAVE

*Note: In the above sequence, '-','**X**','**B**','**Z**','**O**' and blank will be automatically filtered out.*

Custom format. This format enables users to use customized information as description line. A custom format sequence should also be started with ">" symbol. Other rules are all the same as that described in FASTA format. An example sequence in custom format is:

>input your description here and end with newline character

ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRTQIWQKHRTSNDS

Warning: DO NOT use mixed format in one file. This may cause serious problem.

Tips: Custom format is applicable to any file format start with ">" in their description line. You can apply custom format to a FASTA format file, but not vise-versa.

PFSP Options Overview

Initiating Scoring Database

Figure 1-1

PFSP was incorporated with two kinds of scoring system. Dayhoff point accepted mutation (PAM) matrix [1] and blocks substitution matrix (BLOSUM) [2]. Both will be initiated at the startup of PFSP.

PAM matrices are based on global alignments of closely related proteins. The PAM 1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. Scores are derived from a mutation probability matrix where each element gives the probability of the amino acid in column X mutating to the amino acid in row Y after a particular evolutionary time, for example after 1 PAM, or 1% divergence (Figure 1-2). A PAM matrix is specific for a particular evolutionary distance, but may be used to generate matrices for greater evolutionary distances by multiplying it repeatedly by itself. However, at large evolutionary distances the information present in the matrix is essentially degenerated. It is rare that a PAM matrix would be used for an evolutionary distance any greater than 256 PAMs.

Since all PAM scoring matrices at a longer evolutionary distance can be mathematically derived from PAM1, PFSP has a whole set of built-in scoring matrix from PAM1 to PAM400. In PFSP, this PAM scoring database was used to calculate Bayesian exact distance. Also, for consistency, PAM matrix was the only scoring matrix used when evolutionary distance was estimated.

Figure 1-2

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

[top row shows original amino acid; left column shows replacement amino acid].

Mutation probability matrix for the evolutionary distance of 1 PAM (i.e. one accepted point mutation per 100 amino acids). An element of this matrix, [Mij], gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 PAM. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

BLOSUM (BLOcks SUbstitution Matrix) matrices are based on local multiple alignments of more distantly related sequences. For instance, BLOSUM 62, the default matrix in most BLAST-based alignment, is a matrix calculated from comparisons of sequences with no less than 62% identity. Unlike PAM matrices, new BLOSUM matrices are never extrapolated from existing BLOSUM matrices, but are always based on local multiple alignments. So, the BLOSUM 80 matrix would be derived from a set of sequences having 80% sequence identity.

BLOSUM scoring matrix was calculated from ***Blocks.dat*** file shipped with PFSP. However, PFSP does not bother to calculate BLOSUM scoring matrix every time. The pre-calculated relative frequency file ***_ProFunSit_BLOSUM_RltvFreq*** gives a fast solution to scoring matrices from BLOSUM100 to BLOSUM14. You will see later in this guide how to create your own BLOSUM scoring matrix and use customized BLOSUM matrix in your alignment.

Note: All scoring matrices in PFSP are scaled to 10×log[base 10]

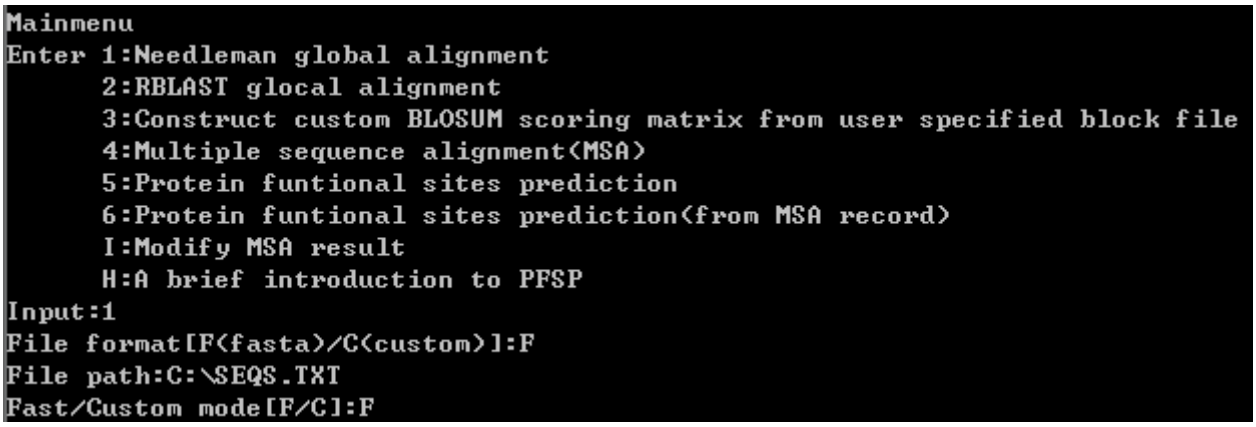
Work Modes

As you may notice in Figure 1-1, PFSP has two work modes available—Fast/Custom mode. Enter F (or f) for fast mode and C (or c) for custom mode (See Figure 3).

If you are not familiar with PFSP, or you have very little idea about how those different algorithms may affect the result, we suggest fast mode. See Table 2-1 for fast mode parameter list.

If you have carefully read this user guide and have a clear mind about which algorithm to use, we strongly suggest custom mode (See Table 2-2). You will be better off if proper algorithms were chosen (especially true for MSA and functional sites prediction).

Figure 2-1



Example: choose fast mode in Needleman global alignment

Table 2-1 Algorithms and parameters used in fast mode

Option	Scoring database	Distance estimation type [b]	Distance estimation subtype	Gap penalty	Word length	Midway iteration anc/amp [c]	Final iteration anc/amp [c]
Needleman global alignment	PAM[D][a]	Approx[A]	Poisson[P]	-3	-	-	-
RBLAST glocal alignment	PAM[D]	Approx[A]	Poisson[P]	-3	3	-	-
Multiple sequence alignment (MSA)	PAM[D]	Approx[A]	Poisson[P]	-3	3	60/3 [d]	60/3 [d]
Protein functional sites prediction	PAM[D]	Approx[A]	Poisson[P]	-3	3	-	-

[a] *Number inside square bracket is the abbreviation for that option*

[b] *Dayhoff PAM scoring rule was forced to be applied when evolutionary distance (PAM) was estimated. BLOSUM scoring rule was available only when using fixed scoring matrix*

[c] *Anchoring point threshold/gap amplification factor*

[d] *Midway/Final iteration are not automatically enabled. PFSP will inquire the user about whether or not enable iteration. Numbers in table are recommended values*

Table 2-2 Available options in custom mode

Option	Scoring database	Distance estimation type [b]	Distance estimation subtype	Gap penalty	Word length	Bayesian step [f]	Gamma	Midway iteration anc/amp [h]	Final iteration anc/amp [h]	Seq number threshold
Needleman global alignment	PAM[D][a] BLOSUM[B]	Approx[A]	Simple[S][c]/Poisson[P]/Gamma[G]	Any [d]	-	-	Any [g]	-	-	-
RBLAST glocal alignment	PAM[D] BLOSUM[B]	Approx[A] Bys Exact[E][e]	Simple[S]/Poisson[P]/Gamma[G] BLAST local[B]/Needleman global[N]	Any	3 or 4	10	Any	-	-	-
Multiple sequence alignment (MSA)	PAM[D]	Approx[A] Bys Exact[E]	Simple[S]/Poisson[P]/Gamma[G] BLAST local[B]/Needleman global[N]	Any	3 or 4	≤10	Any	<=100/ <10	<=100/ <10	-
Protein functional sites prediction	PAM[D]	Approx[A] Bys Exact[E]	Simple[S]/Poisson[P]/Gamma[G] BLAST local[B]/Needleman global[N]	Any	3 or 4	≤10	Any	-	-	≥10

[a] *Number inside square bracket is the abbreviation for that option*

[b] *Dayhoff PAM scoring rule was forced to be applied when evolutionary distance (PAM) was estimated. BLOSUM scoring rule was available only when using fixed scoring matrix*

[c] *Simple PAM distance=(1-(match ×2)/sum of length of two aligned sequences) ×100. Similarity% =(1-PAMdistance[Simple]/100) ×100%*

[d] *We suggest -3 for scoring matrix scaled to 10 ×log[base 10]*

[e] *Bayesian exact distance. Including BLAST local distance and Needleman global distance*

[f] *This term is used to adjust the balance between Bayesian speed and accuracy. Use 10 unless there is a good reason to use a lower value (lower speed also)*

[g] *When a = 2 is used, distance is close to Dayhoff's (1978) [1] PAM distance per site (0.01 PAM)*

[h] *Anchoring point threshold/gap amplification factor*

Needleman Global Alignment

To have a better understanding of this section, the following knowledge is necessary.

PAM scoring matrix

BLOSUM scoring matrix

This algorithm was put forward by Needleman,S.B and Wunsch,C.D. [3] in 1970. This algorithm is a simple and beautiful (but less useful) way to find the maximum match between two sequences. Mathematically, Needleman global alignment was realized by dynamic programming, a small programming trick that solves the main problem without being bothered by overlapping subproblems. You can try this algorithm if the pair of sequences has very high similarity, or if RBLAST glocal rule does not really meet your need. The following example illustrated the steps needed to make pair wise Needleman global alignment in custom mode.

Sample sequences

>gi|159162458|pdb|1I6F|Insect-Specific Neurotoxin Variant 5 (Cse-V5)

KDGYPVDSKGCKLSCVANNYCDNQCKMKKASGGHCYAMSCYCEGLPENAKVSDSATNICG

>gi|158931147|sp|P60213.2|SC49A_TITCA Toxin Tc48b/Tc49a

KDGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAWMACYCYSMPDWVKTWSRSTNRCGR

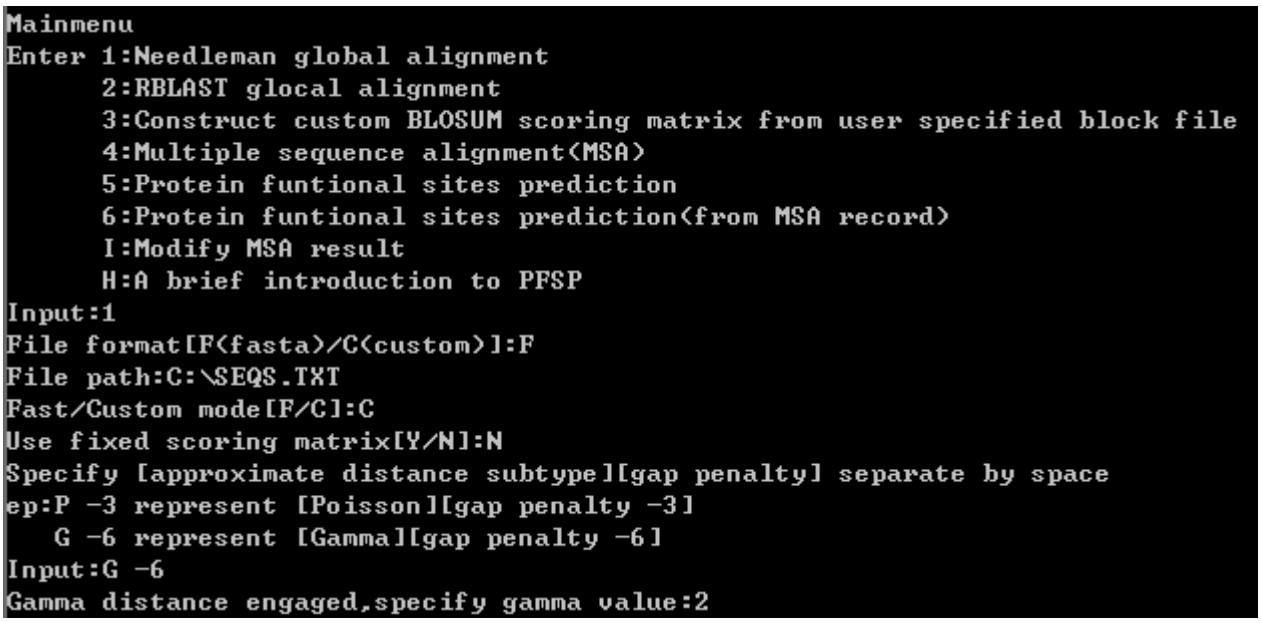
>gi|158705857|sp|P0C5F0.1|KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)

KIDGYPVDNWNCKRICWYNNKYCYDLCKGLKADSGYCWGWTLSCYCEGLPDNARIKRGGRCN

Store above FASTA format sequences in C:\seqs.txt

Figure 3-1

(Needleman global alignment; distance subtype: Gamma; gap penalty: -6; gamma value: 2)



Press enter and the following message appears

Figure 3-2

C:\PFSP\NWALN\NWALN_3_N_D_G_-6.txt means [Needleman-Wunsch global alignment] _ [3 sequences] _ [scoring matrix not fixed] _ [Dayhoff PAM database] _ [Gamma distance] _ [gap penalty -6]

All Needleman global alignment result was saved to C:\PFSP\NWALN directory. The alignment result will be automatically opened in a notepad (Figure 3-3)

Figure 3-3

```
Protein Functional Sites Prediction--Needleman Global Alignment:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 3 sequences   Gap penalty:-6  Score database:DayhoffPAM1-400  Approximate distance:G  Gamma:2

ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
Estimated distance:PAM78  Variance:1.46484  Similarity:51.6129%   Score(10*log[base10]):231.946
          10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+---
KDGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAWMACYCYSMPDWVKTWSRS-TNRCGR
|||| | * ||| *| *    || |*|   |*  | ||| |*||| **|*  |  | | || ||
KDGYPVDSKGCKLSCVANN--YCDNQCKMKKASGGHCYA-MSCYCEGLPENAKV-SDSATNICG-

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
Estimated distance:PAM67  Variance:1.16782  Similarity:55.7377%   Score(10*log[base10]):171.267
          10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+------
KIDGYPVDNWNCKRICWYNNKYCYDL-CKGLKADSGY-CWGWTLSCYCEGLPDNARIKRGGRCN---
| ||||||*  ||  |  || || |  ||  || ||  |   *  |||||||*||**      |
K-DGYPVDSKGCKLSCVANN-YC-DNQCKMKKA-SGGHCYAMS--CYCEGLPENAKVSDSAT-NICG

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
Estimated distance:PAM89  Variance:1.83333  Similarity:47.619%   Score(10*log[base10]):89.819
          10        20        30        40        50        60        70
+---------+---------+---------+---------+---------+---------+---------+
KIDGYPVDNWNCKRICWYNNKYC---YDL--CKGLKA-DSGYCWGWTLSCYCEGLPDNARIK-RGG-RCN-
| ||| | |  ||    ||       |    |  *|* | ||| *|   ||| **||  *   |*  ||
K-DGYLVGNDGCK----YNCLTRPGHYCANECSRVKGKD-GYCYAWMA-CYCYSMPDWVKTWSRSTNRCGR
```

The title shows the date and the alignment parameters. Sequence ID was their order of appearance in the original file. "|"means identical, "" means the two aligned AAs have score greater than 0 in corresponding scoring matrix. Gaps at the terminal of sequences were not scored.*

Since we have arbitrarily chosen gamma distance and arbitrarily set gamma parameter to 2, we are not sure if these parameters are optimal condition for the current sequences. Actually, several gaps with unreasonable length were seen in above result. The length of gap can be reduced by increasing gap penalty, such as minus 10 in this example (Figure 3-3). You may also turn to commonly used fixed scoring matrix PAM250 or BLOSUM62. If none of these choices meet your requirement, use RBLAST instead.

Generally, Needleman rule was not preferred in that it maximizes the matching score on a global range, which may not truly represent the relationship between two sequences. The worst thing happens when the two sequences have very different lengths and one is a block of another. In such cases, Needleman rule forcibly elongates the shorter one by introducing in unreasonably long gaps. Although introducing in gap penalty can to some extent compensate this apparent drawback, it does not provide a fundamental solution to common alignment works. However, when sequence pairs are similar in length and have visually discernible similarity, Needleman global rule will prove similar result as other local/glocal algorithms. Example shown in RBLAST glocal alignment section best illustrated the difference between Needleman global rule and RBLAST glocal rule.

To allow for arbitrary gap length (find the maximum number of match between two sequences) in an alignment, use gap penalty 0. Increase gap penalty results in decrease in gap length. All AAs pile up to the left end of the aligned sequences when gap penalty is very high (such as -30).

Figure 3-3

```
Protein Functional Sites Prediction--Needleman Global Alignment:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 3 sequences   Gap penalty:-10  Score database:DayhoffPAM1-400  Approximate distance:G  Gamma:2

ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
Estimated distance:PAM78  Variance:1.46484  Similarity:51.6129%   Score(10*log[base10]):197.35
          10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+----
KDGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAWMACYCYSMPDWVKTWSRSTNR-CGR
|||| | * ||| *| *    || |*|   |*  | ||| |*||| **|*  |  | |*   ||
KDGYPVDSKGCKLSCVANN--YCDNQCKMKKASGGHCYA-MSCYCEGLPENAKV-SDSATNICG-

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
Estimated distance:PAM67  Variance:1.16782  Similarity:55.7377%   Score(10*log[base10]):156.985
          10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+-----
KIDGYPVDNWNCKRICWYNNKYCYDL-CKGLKADSGYCWGWTLSCYCEGLPDNARIKRGGRCN---
| ||||||*  ||  |  || || |  ||  ||  | |   *  |||||||*||**      |
K-DGYPVDSKGCKLSCVANN-YC-DNQCKMKKASGGHCYAMS--CYCEGLPENAKVSDSAT-NICG

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
Estimated distance:PAM89  Variance:1.83333  Similarity:47.619%   Score(10*log[base10]):99.9628
          10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+------
KIDGYPVDNWNCKRICWYNNK-YCYDLCKGLKA-DSGYCWGWTLSCYCEGLPDNARIK-RGG-RCN-
| ||| | |  ||  |      || * |  *|* |* || *|   ||| **||  *   |*  ||
K-DGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDG-YCYAWMA-CYCYSMPDWVKTWSRSTNRCGR
```

Sometimes, use higher gap penalty may result in higher alignment score. Such as the 3rd alignment in above example, this pair scored 89 when gap penalty is -6 and scored 99 when the penalty is -10. This is because several long gaps were abolished when using stricter penalty scheme.

For more information about Needleman global alignment

http://www2.cs.uh.edu/~zhenzhao/Review/alignment.htm

http://en.wikipedia.org/wiki/Needleman-Wunsch_algorithm

http://en.wikipedia.org/wiki/Dynamic_programming

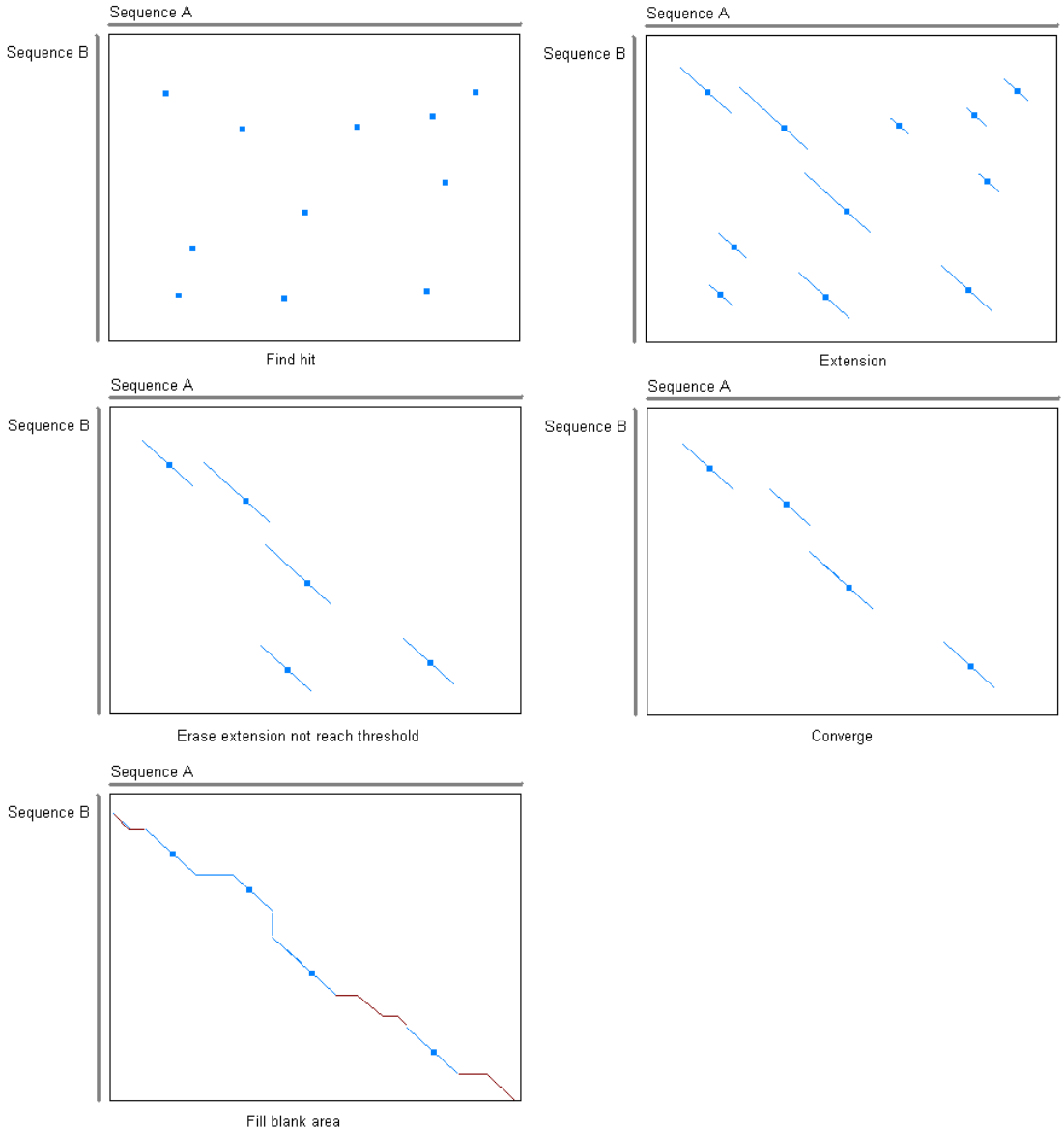
RBLAST Glocal Alignment

To have a better understanding of this section, the following knowledge is necessary.

Genetic distance [Approximate/Bayesian exact]

Basic local alignment searching tool (BLAST) algorithm was put forward by Altschul, S.F. et al [4] in 1990. It has now become one of the most widely used bioinformatics algorithm. The core idea of BLAST is to construct word libraries for both query and target (or database) sequences. Usually the word length is 3 or 4. The first appearance of a word adds a new record storing word and

Figure 4-1



RBLAST glocal alignment work flow. Hit and extension score thresholds were pre-calculated and stored in _ProFunSit_t_s_thld. A new record will be added to this file if new alignment condition is encountered.

location information to the library. If same word record was found in the library, corresponding location repertoire was updated to contain the new location. Alignments of word pairs were then conducted to identify those pairs with matching score exceeding a certain threshold (called a hit). Hit was extended towards both ends until the score

dropped below certain threshold. In PFSP, the threshold for hit was defined as the 95th percentile (a commonly used standard for statistical significance) of scores produced by N (N = 5000) randomly [a] generated word pairs of length W (word length) ([a] *Not really random, but depends on the natural frequency of AAs*). Different from regular BLAST, RBLAST does not control the extension by supervising the alignment score, it rather stops extension when W consecutive negative score (<0) mismatches were encountered. An apparent benefit of this rule is that we have direct control of the maximum number of consecutive negative matches. So, for protein pairs that are distant from each other, a longer W may generate better result than a shorter one. However, longer W may give very same result if protein pairs are evolutionarily close. The threshold used to disqualify extensions was defined as the 95th percentile of maximum extension scores produced by N (N = 5000) randomly generated protein sequence pairs of length S (the length of the shorter sequence in an alignment). So the hit score and extension score threshold vary with different word lengths and scoring matrices. After erasing disqualified extensions, the remaining segments were converged according to their locations and scores. Blank areas not covered in the BLAST step were filled by applying global alignment rule. The whole alignment process was described in figure 4-1.

RBLAST is by no mean the commonly used BLAST that you see elsewhere. Serving as a fundamental module in MSA, RBLAST does not control gap length with gap penalty. It always ensures maximum match between segments that were not covered in the local alignment step. This maximum-global-interspaced-local-alignment rule was proved to be an efficient way to identify evolutionarily identical AAs when a large quantity of sequences is to be aligned. Unreasonable matches will be modified in midway/final MSA iteration while considering the overall alignment situation. The benefit of this scheme is that we do not have to make arbitrary decision on which pair of AAs (distant from block) should be matched when sufficient evidence is not available. One possible disadvantage is that in single pair alignment, gaps with unreasonable length may appear.

The following example illustrated the use of RBLAST glocal alignment in custom mode

Figure 4-2

(RBLAST global alignment; distance type: Approximate; distance subtype: Gamma; gap penalty: -6; word length: 3; Gamma value: 2)

```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      I:Modify MSA result
      H:A brief introduction to PFSP
Input:2
File format[F(fasta)/C(custom)]:F
File path:C:\SEQS.TXT
Fast/Custom mode[F/C]:C
Use fixed scoring matrix[Y/N]:N
Specify [distance estimation type][subtype][gap penalty][word length] separate by space
ep:E B -3 3 represent [Bayesian exact][BLAST local][gap penalty -3][word length 3]
ep:A P -6 3 represent [approximate][Poisson][gap penalty -6][word length 3]
Input:A G -6 3
Gamma distance engaged,specify gamma value:2
RBLAST pairwise alignment in progress...100.00%
Result has been saved to c:\PFSP\RBLAST\RBLAST_3_N_D_A_G_-6.txt
```

C:\PFSP\RBLAST\RBLAST_3_N_D_A_G_-6 means [RBLAST glocal alignment] _ [3 sequences] _ [scoring matrix not fixed] _ [Dayhoff PAM database] _ [Approximate distance] _ [Gamma distance] _ [gap penalty -6]

All RBLAST glocal alignment result was saved to C:\PFSP\RBLAST directory. The alignment result will be automatically opened in a notepad (Figure 4-3)

Figure 4-3

```
Protein Functional Sites Prediction--RBLAST:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 3 sequences   Word length:3  Gap penalty:-6  Score database:DayhoffPAM1-400  Approximate distance:G  Gamma:2

ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
Estimated distance:PAM78  Variance:1.46484  Similarity:51.6129%   Score(10*log[base10]):239.167
          10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+---
KDGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAWMACYCYSMPDWVKTW-SRSTNRCGR
|||| | * ||| *| *   *|| |*|   |*  | ||| |*||| **|*  |   | *|| ||
KDGYPVDSKGCKLSCVA--NNYCDNQCKMKKASGGHCYA-MSCYCEGLPENAKVSDS-ATNICG-

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
Estimated distance:PAM67  Variance:1.16782  Similarity:55.7377%   Score(10*log[base10]):206.966
          10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+---
KIDGYPVDNWNCKRICWYNNKYCYDLCKGLKADSGYCWGWTLSCYCEGLPDNARIKRGGRCN---
| ||||||*  ||  |  || || * ||  ||  | |   **||||||||*||**      |
K-DGYPVDSKGCKLSCVANN-YCDNQCKMKKASGGHCY--AMSCYCEGLPENAKVSDSAT-NICG

ID 2 Gi_number:158705857 Accession:P0C5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)
ID 1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
Estimated distance:PAM89  Variance:1.83333  Similarity:47.619%   Score(10*log[base10]):147.927
          10        20        30        40        50        60
+---------+---------+---------+---------+---------+---------+-----
KIDGYPVDNWNCKRICWYNNK-YCYDLCKGLKADSGYCWGWTLSCYCEGLPDNARIK-RGG-RCN-
| ||| | |  ||  |      || * |  *|*  ||| *| **||| **||  *   |*  ||
K-DGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAW-MACYCYSMPDWVKTWSRSTNRCGR
```

Sequence data was shown in Needleman global alignment section. Similarity is not available in Bayesian exact distance mode. Differing the gap penalty in RBLAST only affects the score, does not affects how sequences were aligned.

In this example, the same distance estimation algorithm (Gamma) and gap penalty was used as in previous chapter. The difference between RBLAST and Needleman global alignment (Figure 3-3) was not very obvious in this example since all sequences have similar length. However, we will see clearly their difference in the following example. Consider:

>father sequence

MKLLLLLTISASMLIEGLVNADGYIRGGDGCKVSCVINHVFCDNECKAAGGSYGYCWAWGLACWCEGLPADREWDYETNTCGGKK

>child block

KAAGGSYGYCWAWGLACWCEGLPA

Child block is actually a segment intercepted from father sequence. Alignment of these two sequences by applying Needleman global rule (set gap penalty to 0 to allow for maximum match) generates the following result

Figure 4-4

```
Protein Functional Sites Prediction--Needleman Global Alignment:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 2 sequences   Gap penalty:0  Score database:DayhoffPAM1-400  Approximate distance:P

ID 1 child block
ID 0 father sequence
Estimated distance:PAM82  Variance:1.1659  Similarity:44.0367%   Score(10*log[base10]):251.236
          10        20        30        40        50        60        70        80
+---------+---------+---------+---------+---------+---------+---------+---------+---
-K--------A---------A-G---G------S------------------YGYCWAWGLACWCEGLPA---------------
 |        |         | |   |      |                  ||||||||||||||||||
MKLLLLLTISASMLIEGLVNADGYIRGGDGCKVSCVINHVFCDNECKAAGGSYGYCWAWGLACWCEGLPADREWDYETNTCGGKK
```

This result obviously is not what we want. You may argue that set gap penalty to -3 or -6 will get rid of those unreasonably long gaps. Let's try -3

Figure 4-5

```
Protein Functional Sites Prediction--Needleman Global Alignment:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 2 sequences   Gap penalty:-3  Score database:DayhoffPAM1-400  Approximate distance:P

ID 1 child block
ID 0 father sequence
Estimated distance:PAM82  Variance:1.1659  Similarity:44.0367%   Score(10*log[base10]):-79.8893
          10        20        30        40        50        60        70        80
+---------+---------+---------+---------+---------+---------+---------+---------+---
-KAAGG-SYGYCWAW-GL--A--C-W----CE-G-L-P-A-------------------------------------------
 |     * *      ||  |    *    |  * *
MKLLLLLTISASMLIEGLVNADGYIRGGDGCKVSCVINHVFCDNECKAAGGSYGYCWAWGLACWCEGLPADREWDYETNTCGGKK
```

This one is even worse. Actually, all AAs in the child block sequence will be pushed into a pile at the left end of the alignment if gap penalty exceeds -8. However, RBLAST glocal rule can do the job excellently

Figure 4-6

```
Protein Functional Sites Prediction--RBLAST:(C++) Copyright 2009,Ruan Xiaoyang
Pairwise align 2 sequences   Word length:3  Gap penalty:-3  Score database:DayhoffPAM1-400  Bayesian exact distance:B  Bayesian step:10

ID 1 child block
ID 0 father sequence
(Local)Estimated distance:PAM3  Variance:0.248197  Score(10*log[base10]):307.912
          10        20        30        40        50        60        70        80
+---------+---------+---------+---------+---------+---------+---------+---------+---
-------------------------------------------KAAGGSYGYCWAWGLACWCEGLPA---------------
                                           ||||||||||||||||||||||||
MKLLLLLTISASMLIEGLVNADGYIRGGDGCKVSCVINHVFCDNECKAAGGSYGYCWAWGLACWCEGLPADREWDYETNTCGGKK
```

Also note that the estimated distance is PAM3 when using Bayesian exact distance (BLAST local) estimation. This is because BLAST local distance rule only take blocks into consideration. See genetic distance for more information about the difference between various kinds of distance estimation methods.

Use fixed scoring matrix

As the hit score and extension score threshold depends highly on the scoring matrix used, arbitrary selection of scoring matrix may result in very bad quality of the alignment result. So, do not use this option unless you have a clear mind about the evolutionary distance between the sequence pairs. Anyway, let PFSP calculates the evolutionary distance (either Approximate or Bayesian exact) is not a bad choice. If you stick to BLOSUM matrix (PFSP distance estimation only gives PAM distance), please use the PAM distance to deduce its BLOSUM equivalent. The following table showed part of the relationship between PAM and BLOSUM.

Table 4-1

PAM	100	120	160	200	250
BLOSUM	90	80	60	52	45

For more information about BLAST and glocal alignment

http://blast.ncbi.nlm.nih.gov/Blast.cgi

http://en.wikipedia.org/wiki/BLAST

Multiple Sequence Alignment (MSA)

To have a better understanding of this section, the following knowledge is necessary.

RBLAST glocal alignment

Genetic distance

Phylogenetic tree

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In general, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Visual depictions of the alignment as in figure 5-2 illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column, and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.

Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments. MSAs require more sophisticated methodologies than pair-wise alignment because they are more computationally complex to produce. Most MSA programs use heuristic methods rather than global optimization because identifying the optimal alignment between more than a few sequences of moderate length is prohibitively computationally expensive. The most commonly used heuristic methods are based on the progressive-alignment strategy [5-7], with ClustalW [8] being the most widely used implementation. The idea is to take an initial, approximate, phylogenetic tree between the sequences and to gradually build up the alignment, following the order in the tree. However, with this method, errors made in the first alignments can not be rectified later as the rest of sequences are added in. The main alternative to progressive alignment is the simultaneous alignment of all the sequences [9;10], but they remain an extremely CPU and memory-intensive approach and are not suitable for massive accurate alignment of thousands of sequences. The T-Coffee method [11] addressed the deterministic feature (once aligned, never modified) of common progressive method by introducing in a weighting term, which takes alignment (both locally and globally) between all pairs of sequences into consideration. This method, as well as its later optimized version (by only weighting sequences that are close enough), suffer from great time consumption when sequences number is considerably large.

In PFSP, MSA is a vital step to protein function sites prediction. As such the quality of MSA directly determined the reliability of functional sites predicted. Since practically hundreds, or even thousands of protein sequences need to be aligned to make the predictions reliable, PFSP incorporated a fast and accurate iteration algorithm that optimize the alignment result in very short time. There is also a panel of algorithms that enable you to balance between accuracy and speed (Table 2). To obtain the fastest speed, use approximate distance->simple distance. To obtain the highest accuracy, use Bayesian exact distance->Needleman global/BLAST local distance. The work flow of MSA in PFSP is as follow.

Import File -> Pair-wise Evolutionary Distance Estimation -> Phylogenetic tree construction -> Sequence alignment (Midway iteration<optional>) -> Final MSA iteration<optional>

The most time-consuming step is pair-wise evolutionary distance estimation, which has a time complexity of $O(n^2)$. The following examples illustrated multiple alignment of 3 sequences and 65 sequences in custom mode (Sample file is available in the package ***SAMPLE_SEQS.txt***).

Figure 5-1

(MSA; distance type: Bayesian exact; distance subtype: BLAST local; gap penalty: -3; word length: 3; Bayesian step: 10)

```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      I:Modify MSA result
      H:A brief introduction to PFSP
Input:4
File format[F(fasta)/C(custom)]:F
File path:C:\SEQS.TXT
Fast/Custom mode[F/C]:C
Specify [distance estimation type][subtype][gap penalty][word length] separate by space
ep:E B -3 3 represent [Bayesian exact][BLAST local][gap penalty -3][word length 3]
ep:A P -6 3 represent [Approximate][Poisson][gap penalty -6][word length 3]
Input:E B -3 3
Specify bayesian exact distance step(large step means higher speed but lower accuracy,recommend 10):10
Creating pair wise bayesian exact distance matrix...100.00%
Constructing NJ tree...
Finding neighbour...
Calculate length...100.00%
Enable midway iterating?[Y/N]:n
Align multiple sequences...100%
Iterate final MSA result?[Y/N]:N
Output result...
MSA Result has been saved to c:\PFSP\MSA\MSA_3_E_B_-3.txt
```

C:\PFSP\MSA\MSA_3_E_B_-3 means [MSA] _ [3 sequences] _ [Bayesian exact distance] _ [BLAST local distance] _ [gap penalty -3]

All MSA result was saved to C:\PFSP\MSA directory. The alignment result will be automatically opened in a notepad (Figure 5-2)

Figure 5-2

```
Protein Functional Sites Prediction--MSA:(C++) Copyright 2009,Ruan Xiaoyang
Multialign 3 sequences  Distype:Bayesian exact distance  Subtype:B  Word length:3  Gap penalty:-3  Score(Gonnet):170.9

Bayesian step:10  Scoring database:DayhoffPAM1-400

ID  0 Gi_number:159162458 Accession:1I6F Locus:Insect-Specific Neurotoxin Variant 5 (Cse-V5)
ID  1 Gi_number:158931147 Accession:P60213.2 Locus:SC49A_TITCA Toxin Tc48b/Tc49a
ID  2 Gi_number:158705857 Accession:POC5F0.1 Locus:KURT1_PARGR Alpha-toxin PgKL1 (Kurtoxin-like 1) (Kurtoxin-like I) (KLI)

Multiple sequence alignment result.'*'stands for identical

            10        20        30        40        50        60
   +---------+---------+---------+---------+---------+---------+-------
   * *** *    **  *      **   *    *     * *      ***   *           *
 0 K-DGYPVDSKGCKLSCVAN-N-YCDNQCK-MKKASG--GHCYA--MSCYCEGLPENAKVSDSATNICG- 0
 1 K-DGYLVGNDGCKYNCLTRPGHYCANECSRV-K--GKDGYCYAW-MACYCYSMPDWVKTWSRSTNRCGR 1
 2 KIDGYPVDNWNCKRICWYN-NKYCYDLCKGL-K--ADSGYCWGWTLSCYCEGLPDNARI--KRGGRCN- 2
   * *** *    **  *      **   *    *     * *      ***   *           *
   +---------+---------+---------+---------+---------+---------+-------
            10        20        30        40        50        60
Pairwise PAM distance matrix (PAM/100)
   ID 0      ID 1      ID 2
    0.000     1.195     1.033
    1.195     0.000     1.933
    1.033     1.933     0.000

Pairwise PAM distance variance matrix (PAM/100)
   ID 0      ID 1      ID 2
    0.000     0.059     0.047
    0.059     0.000     0.154
    0.047     0.154     0.000

Evolutionary tree.Values inside bracket are branch length (PAM/100)
1 (1.048)----(0.885)2
0 (0.148)----(0.000)1 2
```

Multiple alignment of 3 sequences by using Bayesian exact distance. The output result includes PAM distance matrix, distance variance matrix and evolutionary tree. Neighbors were connected by "----". A MSA record file named MSA_3_E_B_-3_record was automatically saved

to C:\PFSP\MSA directory. The record file enables you to do functionally sites prediction directly from option 6 rather than have to align the sequences again (Note: at least 20 sequences are required to enable functional sites prediction).

Now we will use a simple example from T-Coffee [11] to show you how the iteration function in PFSP correct alignment error. Consider the following sequences

>seqa

GARFIELDTHELASTFATCAT

>seqb

GARFIELDTHEFASTCAT

>seqc

GARFIELDTHEVERYFASTCAT

>seqd

THEFATCAT

Figure 5-3

```
           10        20
  +---------+---------+--
                  ** ****
0 GARFIELDTHELAST-FA-TCAT 0
1 GARFIELDTHE-----FASTCAT 1
2 GARFIELDTHEV-ERYFASTCAT 2
3 ------------THE-FA-TCAT 3
                  ** ****
  +---------+---------+--
           10        20
Evolutionary tree.Values inside bracket are branch length (PAM/100)
0 (0.120)----(0.462)3
1 (-0.016)----(0.125)2
0 3 (-0.270)----(0.000)1 2
```

As we have noted in figure 5-3, the red frame marked out an unreasonable mismatch where ‘THE’ of the 3rd sequence should obviously be matched with ‘THE’s of the rest of sequences at position 8, 9, 10. This error arises because ID-0 sequence was first aligned with ID-3 sequence by RBLAST, which considers ‘THEFATCAT’ as a single block (for there is no good reason to insert 5 consecutive gaps and consider ‘THE’ as a separate block at that stage). However, after aligning four sequences together, it becomes obvious that ‘THE’ formed a separate block. This apparent matching error can be corrected by using final MSA iteration, setting anchoring point threshold to 60 and gap amplification factor to 3.

Figure 5-4 *(Enable final MSA iteration)*

```
Input:4
File format[F(fasta)/C(custom)]:C
File path:D:\TEST.TXT
Fast/Custom mode[F/C]:C
Specify [distance estimation type][subtype][gap penalty][word length] separate by space

ep:E B -3 3 represent [Bayesian exact][BLAST local][gap penalty -3][word length 3]
ep:A P -6 3 represent [Approximate][Poisson][gap penalty -6][word length 3]
Input:A G -3 3
Gamma distance engaged,specify gamma value:2
Same distance matrix record found!Use this record file?[Y/N]:Y

Constructing NJ tree...
Finding neighbour...100.00%
Calculate length...100.00%
Enable midway iteration?[Y/N]:N
Align multiple sequences...100%
Iterate final MSA result?[Y/N]:Y
Specify [anchoring point threshold][gap amplification factor] seperate by space
ep:90 3 represent [anchoring point threshold 90%][gap amplification factor 3]
Input:60 3
Iterating...
Are you satisfied with this MSA result?[Y/N]:Y
```

When midway/final MSA iteration was enabled, a dialogue will show asking the user to input anchoring point threshold and gap amplification factor. Then a preview notepad will be automatically opened (Figure 5-5). Another dialogue will ask the user whether he/she is satisfied with the result being shown. If PFSP gets affirmative answer, a formal output will be generated. Also note the dialogue "same distance matrix record found…..". If you are run the same sequence database with same distance estimation algorithm repeatedly, please input 'Y'. This can prevent PFSP from doing overlapping work and can save you A LARGE AMOUNT OF time when there are hundreds and thousands of sequences.

Figure 5-5

```
This MSA result is for your preview only.Please delete this temp file in C:\PFSP\TEMP
Tips:Increase gap amplification factor to reduce gap.Use moderately lower anchor threshold to allow for more anchoring point
Anchor threshold:60  Gap amplification factor:3  Length:22  Length diff:1  Score(Gonnet):37.2167
          10        20
+---------+---------+
         ***     ** ****
--------THE----FA-TCAT
GARFIELDTHELASTFA-TCAT
GARFIELDTHEVERYFASTCAT
GARFIELDTHE----FASTCAT
         ***     ** ****
+---------+---------+
          10        20
```

The wrongly matched 'THE' has now been shifted to the correct location. The item "Length diff" means the length difference of the alignment before and after iteration. Score was calculated as the average pair-wise score by applying Gonnet scoring matrix [12]. (Note: Sequences were ordered according to phylogenetic tree structure in preview mode, whereas in the formal output they were reordered according to their appearance in the original file).

Anchoring point threshold defined the threshold up which a column will be marked as anchoring point. Such as 60, which means any columns have 60% or more identical AA will be identified as anchoring point. The "score" of a position for a specific kind of AA is calculated as

*Score = [number count of same kind of AA] – {[distance to the nearest anchoring point]×[sequence number/10]×[**gap amplification factor**]}.*

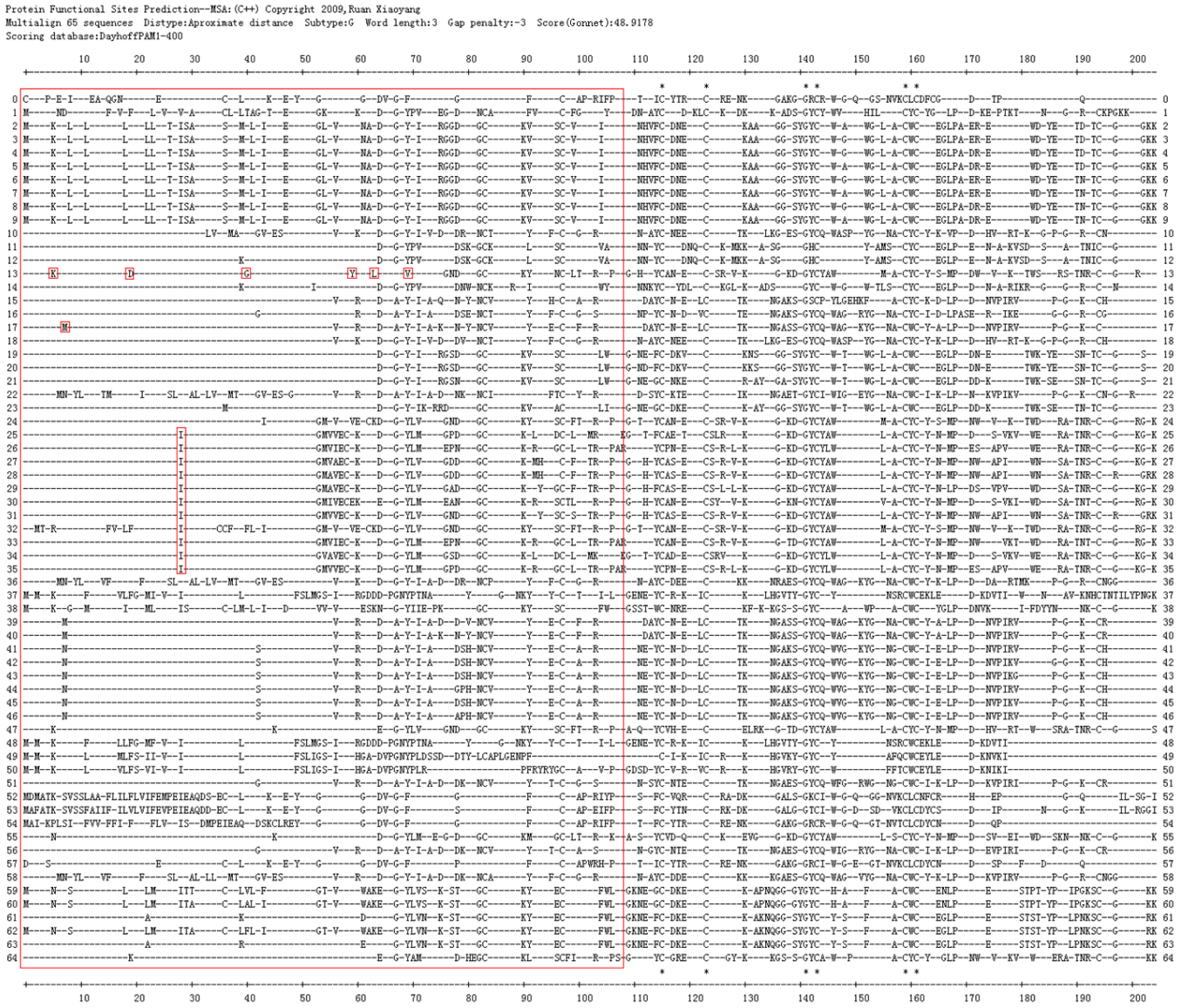
Lowering the threshold will result in more anchoring points. Increase the amplification factor will make the AAs surround closely around anchoring point. When iteration was enabled, the score of AA will be reevaluated throughout the gap area around it. It will be shifted to the position that has the highest score. Generally, depending on actual situation, 2-3 iterations are needed to reach convergence.

Different from **final iteration**, **midway iteration** will iterate the alignment result at each step of the MSA when there are >=3 sequences. Generally, midway iteration is far more effective than final iteration, especially when sequence number is very large. We suggest that users mainly use midway iteration, and use final iteration if further improvement is necessary.

Cooperating with RBLAST, a potential strength of this midway/final iteration scheme is that we do not have make decision on which pair of AAs (that is distant from known block) should be matched when sufficient evidence is not available. It also provides us with the opportunity to correct errors made in previous alignment steps.

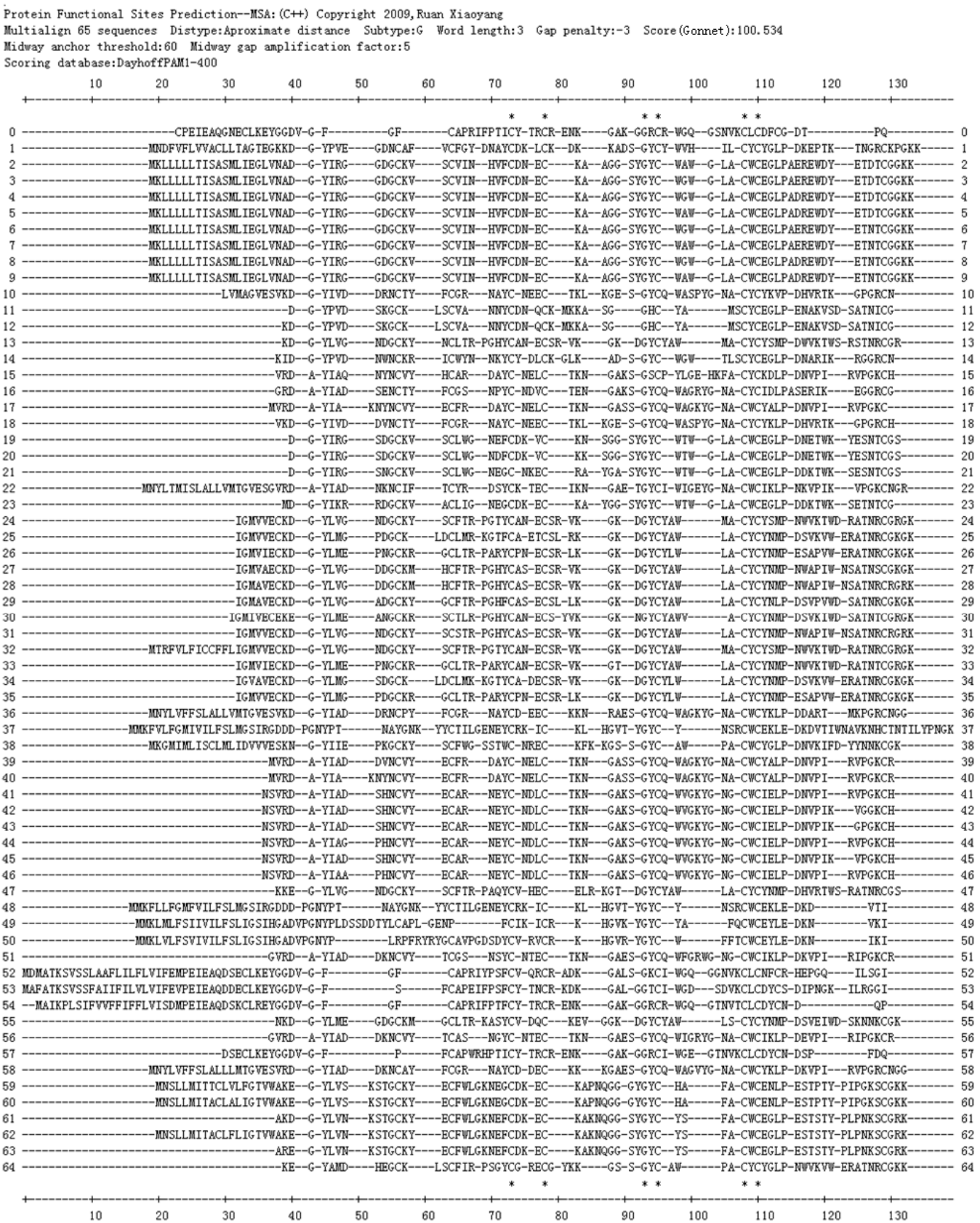
Figure 5-6 and 5-7 illustrated multiple alignment of 65 sequences with and without midway iteration. The improvement is both visually discernable and quantifiable by scoring.

Figure 5-6



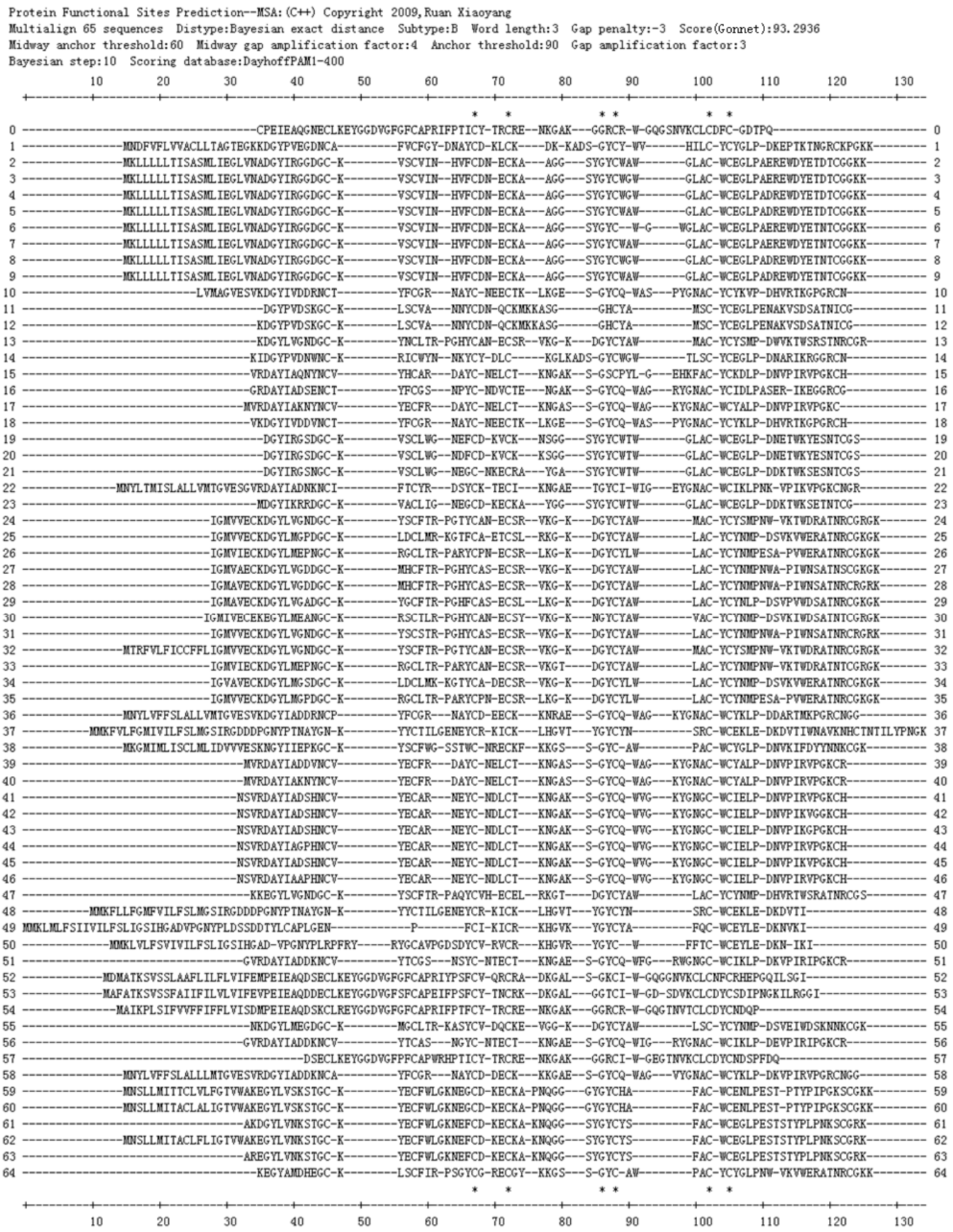
Multiple alignment of 65 sequences without iteration. The total length is 204. Average score is 48.9. The quality of the right half is not so poor but still acceptable. However, a lot of obvious matching errors and unreasonable gaps appeared at the left side of the alignment. We have marked out a small part of these errors.

Figure 5-7



Multiple alignment of 65 sequences with midway iteration. Anchor threshold was set to 60 and gap amplification factor was set to 5. The total length is 139. Average score is 100.5. The errors shown in figure 5-5 were corrected.

Figure 5-8



Apply Bayesian exact distance estimation. Enable midway and final iteration. This alignment result identified same identical AA and has similar score as the result shown in figure 5-7 where Gamma distance estimation was applied.

Actually, we may see even higher score and shorter alignment length if we set the amplification factor higher. And the alignment score will finally converge to a steady point when amplification factor is very high (>10 in most cases, depends on anchor threshold). The increase in score is primarily resulted from decreased gap number. However, this is not equal to say that the quality is also better. Since sometimes gaps are necessary to ensure correct alignment, which is especially true when one sequence was constructed by several separate blocks of another. The users' task is to choose appropriate anchor threshold and amplification factor, and to make both visual and mathematical judgment.

When sequence number is large, different distance estimation algorithms may generate visually discernable different alignment results if iteration was turned off. However, it is likely that the results are very similar when midway iteration was enabled. Figure 5-8 is the alignment result by applying Bayesian exact distance (BLAST local).

In principle, the MSA in PFSP combine the best properties of global and local alignment. It provided a simple, flexible, fast and, most importantly, accurate solution to the problem of multiple alignment of several hundred, or even thousands of protein sequences. All these efforts are made to ensure the functional sites prediction in the next step reliable and reproducible.

Learn more about different kinds of genetic distance estimation.

For more information about multiple sequence alignment

http://en.wikipedia.org/wiki/Multiple_sequence_alignment

http://www.ebi.ac.uk/Tools/clustalw2/index.html

http://bibiserv.techfak.uni-bielefeld.de/dca/

http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html

Protein Functional Sites Prediction

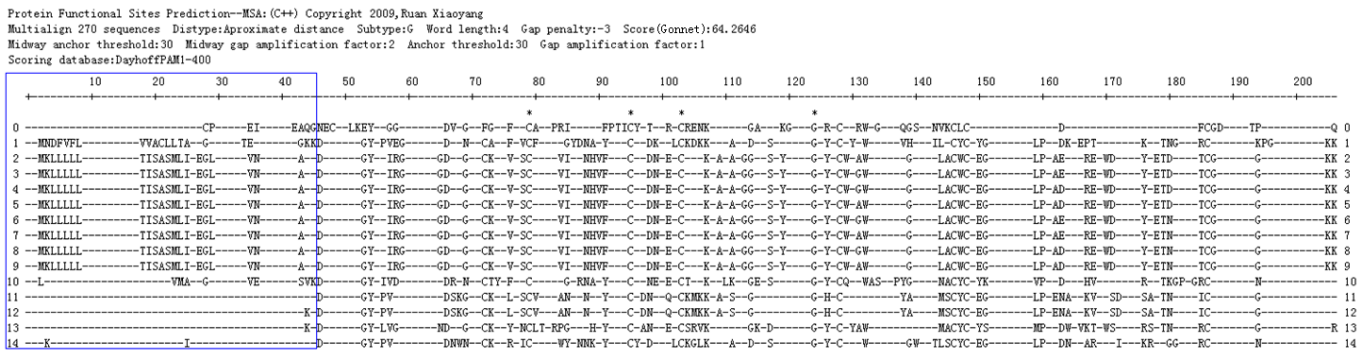
To have a better understanding of this section, the following knowledge is necessary.

Multiple sequence alignment

The realization of functional sites prediction primarily depends on the coupling energy fluctuation bring about by mutation at specific location. The idea of statistical coupling energy has been systematical described by Steve W. Lockless, *et al*.[13]. PFSP added to their analyses an average coupling energy calculated by averaging over the coupling energy changes at the rest of sites when one site mutated. PFSP also output the estimated coupling energy for every site in the most direct way. These results could be used as aids to identify potential functional sites and help biologists to optimize their mutation plan (as long as enough sequence is available to justify the statistics).

There are two ways to do functional sites prediction in PFSP. One is from raw sequences and another is from MSA record (located in C:\PFSP\MSA, or other valid directory if you are using ClustalW MSA result). In the next example 270 sequences will be aligned by applying Gamma distance estimation algorithm. To illustrate how to start from MSA record, we first align the sequences from option 4. The following result was obtained. Meanwhile, a record file named *MSA_270_A_G_-3_RECORD* was generated.

Figure 6-1



Multiple alignment of 270 sequences by using Gamma distance estimation. In this step, we set midway anchor threshold to 30 (as the sequences vary widely in length, only universally shared segments can be properly aligned if high threshold was used. You may have a better view of this by looking at Figure 5-7 and 5-8). This can help us to get a better alignment for sites from 0 to 45. In the next step, a higher threshold will be used to ensure a better alignment of the common areas (>=46).

Figure 6-2

```
Mainmenu
Enter 1:Needleman global alignment
      2:RBLAST glocal alignment
      3:Construct custom BLOSUM scoring matrix from user specified block file
      4:Multiple sequence alignment(MSA)
      5:Protein funtional sites prediction
      6:Protein funtional sites prediction(from MSA record)
      I:Modify MSA result
      H:A brief introduction to PFSP
Input:6
File path of MSA result:C:\PFSP\MSA\MSA_270_A_G_-3_RECORD
Specify sequence number threshold(>=10):10
Specify [MSA result format]
ep:P represent [PFSP MSA result]
   C represent [ClustalW MSA result]
Input:P
Import file...
Estimate conservative sites...
Compute coupling energy...100.00%
Compute coupling energy threshold...100.00%
Output result...
PFSP Result has been saved to C:\PFSP\FUNCSITE\PFSP_270_10.txt
```

We then select option 6, which enables us to perform functional sites prediction from MSA record (Figure 6-2).

Two MSA formats are supported in PFSP.

1 The *MSA_XXX_RECORD* files generated by PFSP. Locate these files in the C:\PFSP\MSA\ folder.

2 The *XXX.aln* files generated by ClustalW. The files should have the following format (Figure 6-3) in order to be properly recognized. If you have no idea whether the format is correct or not, the best choice is leave it intact.

Figure 6-3

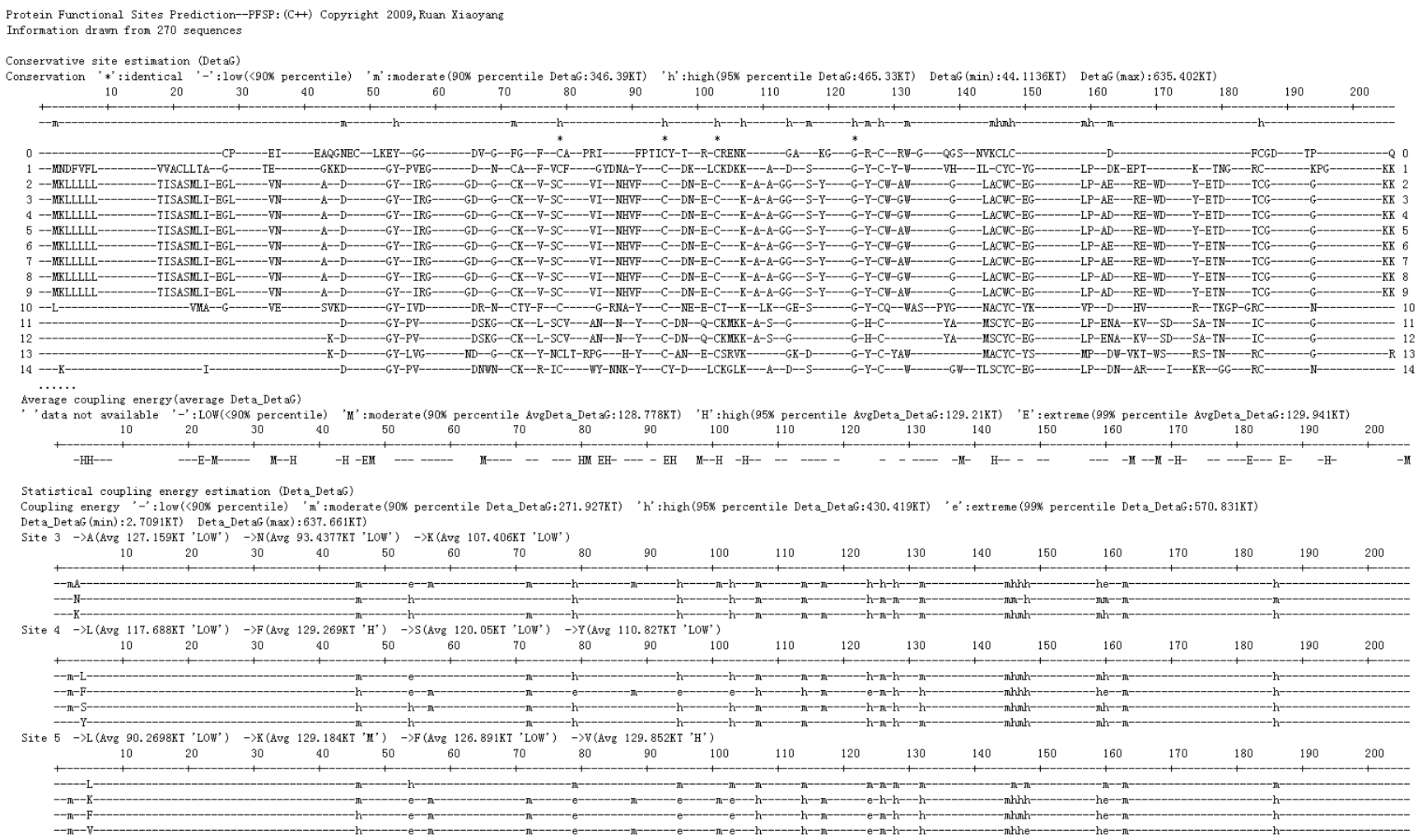
```
CLUSTAL 2.0.10 multiple sequence alignment


gi|116270728|sp|Q1I180.1|SCX1_      ----------------------IGMVIEC-KDGYLM-EPNGCKR-GCLTR
gi|116256066|sp|Q1I176.1|SCX10      ----------------------IGMVVEC-KDGYLM-GPDGCKR-GCLTR
gi|116256068|sp|Q1I172.1|SCX12      ----------------------IGMVIEC-KDGYLM-EPNGCKR-GCLTR
gi|116256072|sp|Q1I167.1|SCX6_      ----------------------IGMIVECEKEGYLM-EANGCKR-SCTLR
gi|116270731|sp|Q1I174.1|SCX4_      ----------------------IGMVVEC-KDGYLV-GNDGCKY-SCFTR
gi|116256069|sp|Q2NME3.1|SCX1_      ---------MTRFVLFICCFFLIGMVVEC-KDGYLV-GNDGCKY-SCFTR
gi|158931147|sp|P60213.2|SC49A      ----------------------------KDGYLV-GNDGCKY-NCLTR
gi|108860962|sp|P84631.1|SCX2_      ----------------------------K-KEGYLV-GNDGCKY-SCFTR
gi|116256074|sp|Q1I163.1|SCX8_      ----------------------IGMAVEC-KDGYLV-GDDGCKM-HCFTR
                                                                                    *

gi|116270728|sp|Q1I180.1|SCX1_      P-ARYCPNECSR--LKGKDGYCYLWLA----CYCYNMPES-APVWERATN
gi|116256066|sp|Q1I176.1|SCX10      P-ARYCPNECSR--LKGKDGYCYLWLA----CYCYNMPES-APVWERATN
gi|116256068|sp|Q1I172.1|SCX12      P-ARYCANECSR--VKGTDGYCYAWLA----CYCYNMPNW-VKTWDRATN
gi|116256072|sp|Q1I167.1|SCX6_      P-GHYCANECSY--VKGKNGYCYAWVA----CYCYNMPDS-VKIWDSATN
gi|116270731|sp|Q1I174.1|SCX4_      P-GTYCANECSR--VKGKDGYCYAWMA----CYCYSMPNW-VKTWDRATN
gi|116256069|sp|Q2NME3.1|SCX1_      P-GTYCANECSR--VKGKDGYCYAWMA----CYCYSMPNW-VKTWDRATN
gi|158931147|sp|P60213.2|SC49A      P-GHYCANECSR--VKGKDGYCYAWMA----CYCYSMPDW-VKTWSRSTN
gi|108860962|sp|P84631.1|SCX2_      P-AQYCVHECEL--RKGTDGYCYAWLA----CYCYNMPDH-VRTWSRATN
gi|116256074|sp|Q1I163.1|SCX8_      P-GHYCASECSR--VKGKDGYCYAWLA----CYCYNMPNW-APIWNSATN
                                        *   *         * *         * *
```

Warning: The program may get crashed if you accidentally altered the contents in the MSA file

Sequence number threshold defined the minimum sequence number difference before and after mutation (too little difference results in overestimation/underestimation of the actual effect). Generally, higher threshold means better statistics but may also results in loss of sites with rare mutations. We recommend 10 as default value. However, if more sequences (such as 1000) are available, a higher threshold could be used.

From figure 6-4 (or sample result in ***SAMPLE_PFSP_270_10 Threshold 30***), we can identify sites with different conservative levels as well as an estimated average coupling energy change for mutations at each site. These results enable you to estimate the global effect of specific mutation on the whole protein sequence. According to result ***SAMPLE_PFSP_270_10 Threshold 30,*** we can identify sites at *22 to V, **47 to S,** 83 to A, 93 to S, 187 to D* that have extreme (>99th percentile) values, and *4 to F, 5 to V, 36 to P, 44 to G, **80 to M,** 84 to K, **94 to L, 101 to R, 105 to S, 143 to N,** 171 to S, 194 to P* that have high (>95th percentile) values, and *24 to F, 33 to S, 48 to E, 65 to N, 81 to I, 98 to Y, 138 to Q, 164 to N, 168 to R, 206 to F* that have moderately high values (>90th percentile). Also note that sites after position 46 are less reliable in this scheme, which left us with sites **4 to F ^{h}, 5 to V^{h}, 22 to V ^{e}, 24 to F ^{m}, 33 to S ^{m}, 36 to P ^{h}, 44 to G ^{h}** that seem to be important.

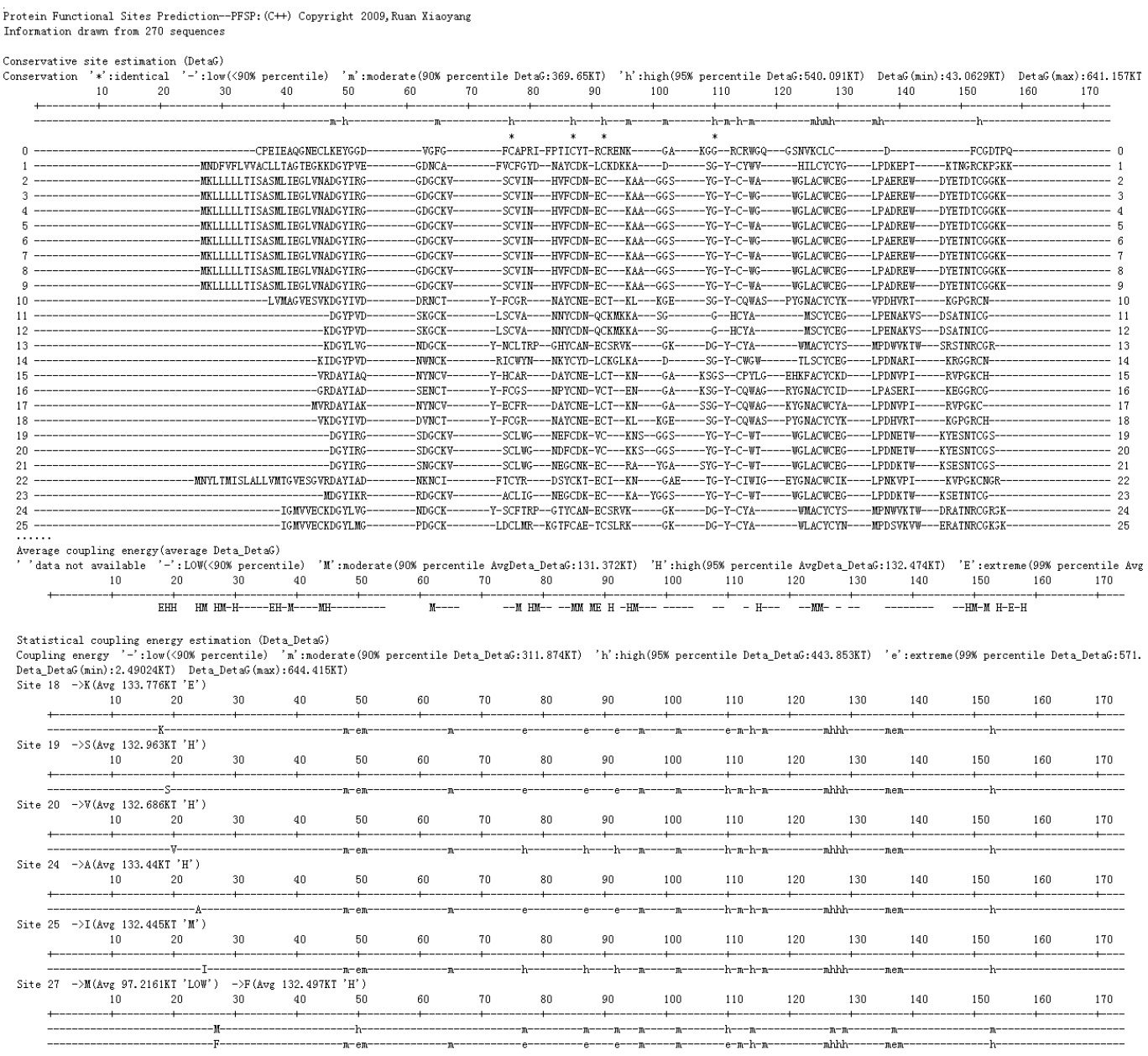
Figure 6-4



Protein functional sites prediction results. It is composed of three parts. Part 1 (Conservative site estimation) showed three conservative levels labeled by '-', 'm', 'h' (for low, medium, high conservation). Part 2 (Average coupling energy) showed the averaged coupling energy changes at the remaining sites when one site mutated. Blank area means no mutation data is available for that site. The labels 'M' 'H' 'E' represented the highest average coupling energy observed at that site. Part3 (Statistical coupling energy) presented the mutation data, if any, for each site. For instance, Site 5 has four mutations in its history (L, K, F, V) and the mutation to V has significant impact on other sites (Avg 129.852KT 'H').

Next we set midway anchor threshold to 60, and refine the final result with anchor threshold 80 and gap amplification factor 4. Figure 6-4 (or ***SAMPLE_PFSP_270_10 threshold 60***) shows part of the alignment result when using these standards. We identified sites at *18 to K, 36 to K, 89 to Q, 156 to T* that have extreme values, and *19 to S, 20 to V, 24 to A, 27 to F, 30 to V, 37 to P, 45 to I,* ***78 to M, 91 to R, 94 to S,*** *115 to L, 149 to T, 154 to K, 158 to F* that have high values, and *25 to I, 28 to I, 39 to A,* ***44 to S,*** *62 to H, 76 to Y, 79 to A, 85 to N,* ***86 to L,*** *88 to R, 95 to V,* ***124 to N,*** *125 to I, 150 to K, 152 to F* that have moderately high values. Also note that sites before position 47 (Figure 6-4, falls outside of the first block) are less reliable in this scheme, which left us with sites **62 to H [m], 76 to Y [m], 78 to M [h], 79 to A [m], 85 to N [m], 86 to L [m], 88 to R [m], 89 to Q [e], 91 to R [h], 94 to S [h], 95 to V [m], 115 to L [h], 124 to N [m], 125 to I [m], 149 to T [h], 150 to K [m], 152 to F [m], 154 to K [h], 156 to T [e], 158 to F [h]** that seem to be important. Brown colored words are duplicate records that appeared in both schemes (Table 6-1).

Figure 6-5



Part of the alignment result by using midway anchor threshold 60. The final result was refined by using anchor threshold 80 and gap amplification factor 4 (as this standard gives the highest score). Different from the result shown in Figure 6-4, this result is more compact and has a clear view of blocks.

Table 6-1

Low Threshold	47 to S [e]	80 to M [h]	94 to L [h]	101 to R [h]	105 to S [h]	143 to N [h]
High Threshold	**44 to S [m]**	**78 to M [h]**	**86 to L [m]**	**91 to R [h]**	**94 to S [h]**	**124 to N [m]**

Sites shown in Table 6-1 are very likely to be important functional sites. First, they were cross-validated in different alignment schemes. Second, their locations are very close to conservative blocks (Figure 6-5). However, this is not to say that the predictions on other sites are not correct. Actually, you can expect to get the same prediction results when sequences are properly aligned.

When limited number of sequence is available, it is always a good idea to do cross-validation in different alignment schemes. This can potentially avoid significances arising from error. When a large number of homologous sequences, say 1000, are available, you can randomly split it into two or three parts and do cross-validation between them. Also, DO NOT merely pay attention to sites with extreme (>99th percentile) value, there isn't much difference between >95th and >99th considering the errors introduced at the sequence selection step and the following parameter selection steps, especially when sequence number is small.

Genetic Distance

To have a better understanding of this section, the following knowledge might be helpful.

PAM distance

There are two ways to calculate genetic distance -> approximate method and exact method. **Approximate method** gives a quick estimation on the evolutionary distance by simply calculating the proportion of difference AAs between two aligned protein sequences. Additional correction algorithm was used to minimize the gap between approximate and exact method.

Simple p-distance

As noted above, the p-distance is merely the proportion of different amino acids between two sequences compared. It is used without additional correction step.

$$p = n_d/n$$

$$V(p) = p(1 - p)/n$$

Here n_d and n are the number of amino acid differences and the total number of amino acids compared, respectively.

Poisson-correction distance

This distance is for estimating the number of amino acid substitutions per site under the assumption that the number of amino acid substitutions at each site follows the Poisson distribution. This estimator (a) and its variance are given by

$$d = -\log_e(1 - p)$$

$$V(d) = p/[(1 - p)n]$$

where *p* is estimated by equation given in simple p-distance.

Gamma distance

This distance is an estimate of the number of amino acid substitutions per site under the assumption that the rate of amino acid substitution varies from site to site and follows the gamma distribution with parameter *a*. This distance and its variance can easily be computed from Nei *et al.*'s (1995) [14] work.

$$d = a[(1 - p)^{-1/a} - 1]$$

$$V(d) = p[(1 - p)^{-(1 + 2/a)}]/n$$

When *a* = 2 is used, *d* is close to Dayhoff's (1978) [1] PAM distance per site (0.01 PAM)

Bayesian exact distance calculates the probability score distribution from PAM 1 to PAM 400 (or higher) [15]. Then a relative probability was calculated to make all probability scores sum to 1. The evolutionary distance (in PAM/100) was obtained by computing the expectation value *E(X)*

$$\boldsymbol{d = E(X) = \sum_{pam=1}^{400} S_{pam}\, P_{pam}}$$

$$\boldsymbol{V(d) = E(X^2) - E^2(X)}$$

Where $\boldsymbol{S_{pam}}$ is the probability score at distance ***pam*** and $\boldsymbol{P_{pam}}$ is the relative probability. Figure 19 illustrated the relative

p distribution from PAM 1 to PAM 270 (Bayesian, BLAST local rule) by aligning the following 2 sequences, whereas Figure 20 used Bayesian Needleman global rule.

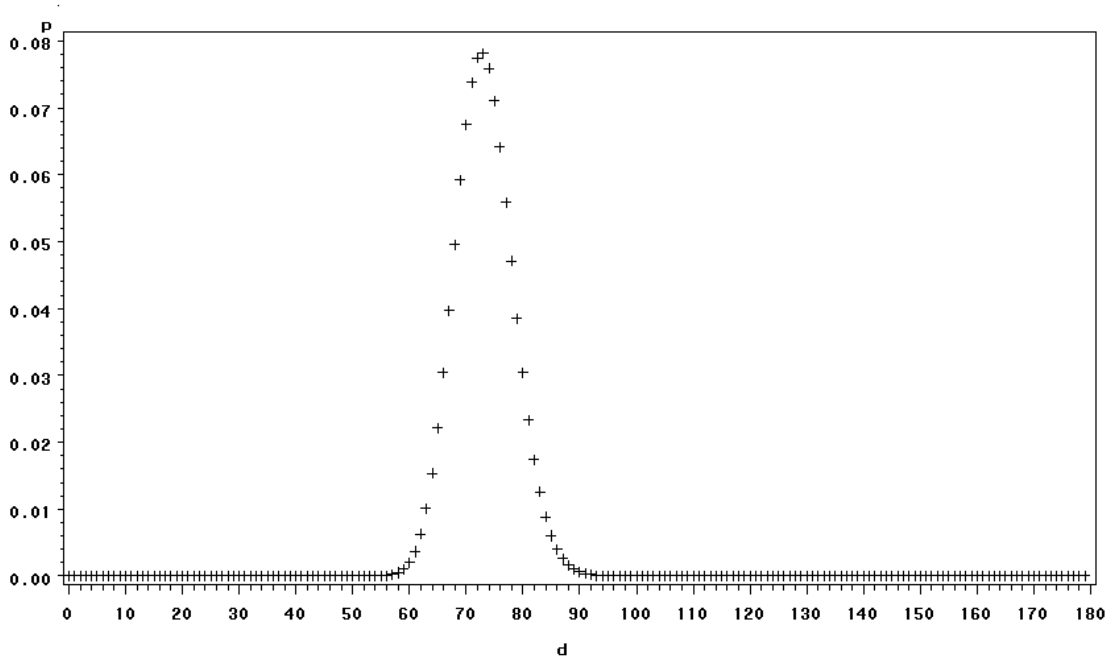
>gi|159162458|pdb|1I6F|Insect-Specific Neurotoxin Variant 5 (Cse-V5)

KDGYPVDSKGCKLSCVANNYCDNQCKMKKASGGHCYAMSCYCEGLPENAKVSDSATNICG

>gi|158931147|sp|P60213.2|SC49A_TITCA Toxin Tc48b/Tc49a

KDGYLVGNDGCKYNCLTRPGHYCANECSRVKGKDGYCYAWMACYCYSMPDWVKTWSRSTNRCGR

Figure 19



In table 3, Gamma algorithm with a=2 gives very similar result as Bayesian exact (Needleman global). We noticed the highest distance when using BLAST local rule. This is because the rest of algorithms maximize the alignment score on a global range, whereas BLAST local rule only consider segments with extension scores above significance threshold. However, different distance estimation methods will give same results if two sequences are very similar.

Figure 20

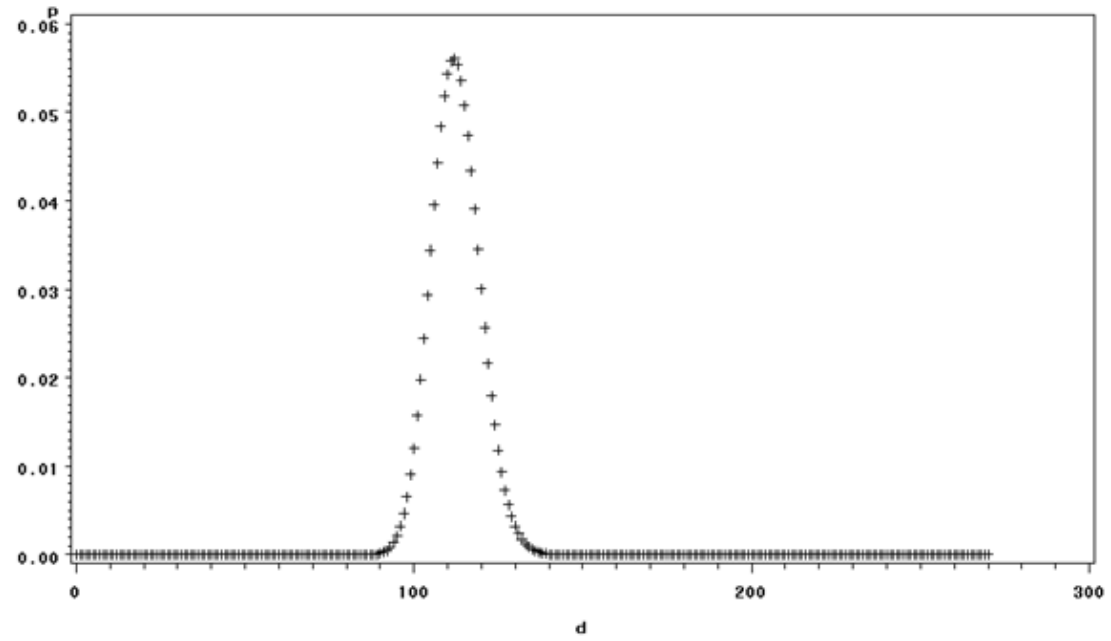


Table 3

Distance estimation type	Distance (PAM)	Variance (PAM)
Simple P-distance	48	0.201
Poisson correlation distance	66	0.756
Gamma distance (a=2)	78	1.465
Bayesian exact (Needleman global)	78	2.917
Bayesian exact (BLAST local)	119	5.878

The ideal way to calculate Bayesian exact distance is try the scoring matrices one by one (from PAM 1 to 400 or higher) on the alignment sequences and identify the one maximizes the alignment score. However, this approach requires prohibitively long time and is not practical when pair wise distance need to be calculated for several hundreds of sequences. In PFSP, when Bayesian exact algorithm was selected, Gonnet scoring matrix [12] was used to guide the initial alignment. Then PAM matrices were then tried on the alignment result to identify the one maximizing the score. Gonnet scoring matrix was derived from an exhaustive matching of the entire MIPS (Munich Information Center for Protein Sequences) database and was recommended as the initial scoring matrix before subsequent refinement was applied.

Bayesian step controlled how PAM matrices were applied to the alignment result. A Bayesian step with value *N* means PAM 1, 1+*N*, 1+2*N*… will be tried in order. Generally, *N*=10 have very similar distance and variance results with *N*=1.

Phylogenetic Tree

A phylogenetic tree or evolutionary tree is a tree showing the evolutionary relationships among various biological species or other entities that are believed to have a common ancestor. In a phylogenetic tree, each node with descendants represents the most recent common ancestor of the descendants, and the edge lengths in some trees correspond to time estimates. Each node is called a taxonomic unit. Internal nodes are generally called hypothetical taxonomic units (HTUs) as they cannot be directly observed. Generally, there are two types of Phylogenetic tree: rooted and un-rooted. A rooted phylogenetic tree is a directed tree with a unique node represents the most recent common ancestor of all the terminals of the tree. Un-rooted trees do not need to make assumption about a common ancestry. Both rooted and unrooted phylogenetic trees can be either bifurcating or multifurcating. A bifurcating tree has a maximum of two descendents arising from each interior node, while a multifurcating tree may have more than two. Current version of PFSP calculates un-rooted tree from pair wise distance matrix by using the neighbor-joining (NJ) method developed by Saitou,N et al. [16] and later modified by Studier,J.A. et al [17]. The result was then used to guide the multiple sequence alignment. An apparent advantage of NJ method lies in its ability to correctly reconstruct the tree from distance matrix D given that the distances in D correspond exactly to those in an actual tree. This property of NJ makes it the best choice to compute phylogenetic tree for large number of homologous proteins.

To construct NJ tree, the pair of sequences which minimizes the total distance between all nodes in the distance matrix was marked as neighbor. This is equal to find the pair of sequences that shares the longest path to the remaining nodes, which can be mathematically attained by minimizing the following term

$$S_{ij} = (N-2)D_{ij} - R_i - R_j$$

Where D_{ij} is the distance between node *i* and *j*. R_i is the sum of distance from node i to all remaining nodes. *N* is the number of nodes.

$$R_i = \sum_{k \neq i} D_{ik}$$

This pair of sequences was then merged into a new node *u*. The distance between *u* and other nodes can be calculated by

$$D_{iu} = \frac{1}{2}\left(D_{ik} + D_{jk} - D_{ij}\right) \text{ for } k \neq i, j$$

The branch lengths from *i* to *u* and *j* to *u* are

$$D_{iu} = \frac{1}{2(N-2)}\left[(N-2)D_{ij} + R_i - R_j\right]$$

And

$$D_{ju} = \frac{1}{2(N-2)}\left[(N-2)D_{ij} + R_j - R_i\right]$$

NJ method has a time complexity of *O* (n^2) to find the neighbor and *O (n)* to update the R_i array. In PFSP, the NJ tree calculation module has been fine-tuned to handle several hundreds of sequences within few seconds.

The NJ result outputted by current version of PFSP has the following appearance

Figure 21

```
Evolutionary tree. Values inside bracket are branch length (PAM/100)
11 (0.012)----(-0.004)12
10 (0.062)----(0.046)18
15 (0.186)----(0.081)17
15 17 (0.040)----(0.124)10 18
15 17 10 18 (-0.160)----(0.260)16
11 12 (0.289)----(0.298)14
11 12 14 (-0.274)----(0.388)13
15 17 10 18 16 (-0.003)----(0.572)0
15 17 10 18 16 0 (-0.312)----(-0.149)11 12 14 13
15 17 10 18 16 0 11 12 14 13 (-0.017)----(0.341)1
2 (0.006)----(0.006)3
6 (0.006)----(0.006)7
6 7 (-0.003)----(-0.003)2 3
4 (0.006)----(0.006)5
8 (0.006)----(0.006)9
6 7 2 3 (0.002)----(0.197)15 17 10 18 16 0 11 12 14 13 1
8 9 (-0.003)----(-0.002)4 5
6 7 2 3 15 17 10 18 16 0 11 12 14 13 1 (-0.187)----(0.000)8 9 4 5
```

Sequences were represented by their corresponding ID in the original file. Negative lengths can be simply replaced with zero.

Construct Custom BLOSUM Scoring Matrix

BLOSUM (BLOcks of amino acid Substitution Matrix) is a scoring matrix used for protein sequence alignment. Different from PAM scoring matrix, BLOSUM is based on local alignments. The idea of BLOSUM was put forward by Henikoff et al [2] in 1992. They scanned the BLOCKS database for very conserved regions of protein families (without gap) and then counted the relative frequencies of AAs and their substitution probabilities. Then, they calculated a log-odds for each of the 210 possible substitutions of the 20 standard AAs. All BLOSUM are based on observed alignments.

With one initial block database file, several sets of BLOSUM matrices can be derived according to different re-clustering levels. This level determined the similarity threshold above which two or more sequences will be merged into one single sequence and then comparing those sequences (that were all more divergent than the given level) only; thus reducing the contribution of closely related sequences. As such BLOSUM80 is used for less divergent alignments, whereas BLOSUM45 is used for more divergent one. This is just the opposite of PAM.

In PFSP, BLOSUM database (from BLOSUM100 to BLOSUM14) derived from standard block file (blocks-5.0.dat. Available at http://blocks.fhcrc.org/blocks/uploads/blosum/) has been pre-calculated and is ready for use. The original block file collected information from 2106 blocks and is competent for alignment work under most circumstances. However, users are allowed to construct their own BLOSUM matrix and use the custom scoring matrix in alignment.

A block file should have the following format.

Figure 22

```
          Blocks Database Version 5.0, June 1992
 ID   GLU_CARBOXYLATION; BLOCK
 AC   BL00011; distance from previous block=(1,64)
 DE   Vitamin K-dependent carboxylation domain proteins.
 BL   ECA motif; width=40; 99.5%=703; strength=2331
 FA10_BOVIN  (     45)  LEEVKQGNLERECLEEACSLEEAREVFEDAEQTDEFWSKY

 FA10_CHICK  (     45)  LEEMKQGNIERECNEERCSKEEAREAFEDNEKTEEFWNIY

 FA10_HUMAN  (     45)  LEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKY

  FA7_BOVIN  (      5)  LEELLPGSLERECREELCSFEEAHEIFRNEERTRQFWVSY
//
```

File format of block. New block file can be downloaded from http://blocks.fhcrc.org/blocks/uploads/blosum/

There should be a title describes the version of the block file. The first word of the title must be “Blocks” (case sensitive). Such as

Blocks hello

Blocks 3000 block

Blocks custom version 1

are all allowed. But

blocks hello

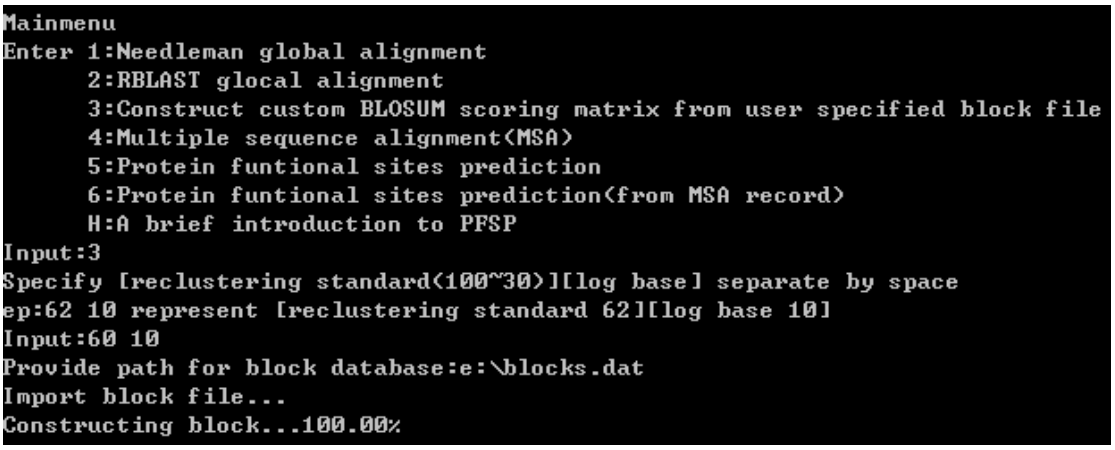
My block 3000

are not allowed. (*Note: PFSP will warn about the use of wrong title, you can choose to proceed if you insist to use wrong title*)

Each block should have ID, AC, DE, BL information in separated line. “//” was used to mark the end of a block. (*Note: Block file information is case and newline sensitive but blank insensitive*)

To create custom BLOSUM matrix, choose option 3, input re-clustering standard and log base value. Then provide the path to custom block file. Here is a sample output of custom BLOSUM60 matrix named with “Blocks hello my new BLOSUM”

Figure 23



The output of BLOSUM file was shown in figure 24

Figure 24

```
Protein Functional Sites Prediction--BLOSUM:(C++) Copyright 2009,Ruan Xiaoyang
Blocks hello my new BLOSUM
Re-clustering percentage 60%

1947 blocks processed  1482 blocks contributed pairs to matrix
568221 total pairs  25967 total sequences  69754 total columns  888228 total AAs
AA frequencies
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
0.077     0.052     0.043     0.051     0.023     0.035     0.055     0.077     0.026     0.066     0.097     0.058     0.025     0.047     0.

Number of pairs count
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
13207.830  2733.636  2339.078  2491.714  1681.164  2394.951  3676.471  6761.970  1288.307  3682.947  5113.048  3701.625  1679.795  1900.902  2
 2733.636  9735.758  2232.794  1822.617   505.430  2844.817  3011.339  2132.458  1569.637  1599.268  2759.618  6800.483   847.344  1042.013  1
 2339.078  2232.794  7093.218  4059.097   496.259  1738.592  2492.663  3183.764  1553.579  1164.774  1754.358  2814.480   710.374   970.777  1
 ......
Frequency
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
    0.023     0.005     0.004     0.004     0.003     0.004     0.006     0.012     0.002     0.006     0.009     0.007     0.003     0.003
    0.000     0.017     0.004     0.003     0.001     0.005     0.005     0.004     0.003     0.003     0.005     0.012     0.001     0.002
    0.000     0.000     0.012     0.007     0.001     0.003     0.004     0.006     0.003     0.002     0.003     0.005     0.001     0.002
  ......
Relative odds
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
    3.876     0.603     0.615     0.558     0.818     0.782     0.759     1.003     0.565     0.634     0.597     0.730     0.779     0.458
    0.000     6.447     0.882     0.613     0.370     1.394     0.934     0.475     1.033     0.413     0.484     2.015     0.590     0.378
    0.000     0.000     6.692     1.629     0.433     1.017     0.922     0.847     1.221     0.359     0.367     0.995     0.591     0.420
  ......
Sij_log 10
A         R         N         D         C         Q         E         G         H         I         L         K         M         F         P
    5.884    -2.200    -2.108    -2.536    -0.871    -1.071    -1.199     0.013    -2.483    -1.980    -2.239    -1.366    -1.086    -3.387
   -2.200     8.094    -0.543    -2.127    -4.323     1.444    -0.298    -3.231     0.142    -3.836    -3.150     3.043    -2.291    -4.231
   -2.108    -0.543     8.256     2.119    -3.634     0.074    -0.351    -0.722     0.866    -4.444    -4.349    -0.020    -2.288    -3.769
  ......
```

To use custom BLOSUM matrix in alignment, just replace the original block file ***Blocks.dat*** in C:\WINDOWS with your own block file (rename to **Blocks.dat**). You can also restore the original BLOSUM matrix by replacing the custom block file with the original one.

Acknowledgement

I would like to express my special gratitude to Dr. Zhu Jun who gave me the ideas and materials. Her help makes it possible for me to start the work on PFSP.

References

1. **Dayhoff MO, R.V.Eck, C.M.Park. A model of evolutionary change in proteins. Atlas of protein sequence and structure 1972; 5:89-99.**

2. **Henikoff S, Henikoff,JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992; 89:10915-9.**

3. **Needleman SB, Wunsch,CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970; 48:443-53.**

4. **Altschul SF, Gish,W, Miller,W, Myers,EW, Lipman,DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10.**

5. **Feng DF, Doolittle,RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 1987; 25:351-60.**

6. **Hogeweg P, Hesper,B. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J Mol Evol 1984; 20:175-86.**

7. **Taylor WR. A flexible method to align large numbers of biological sequences. J Mol Evol 1988; 28:161-9.**

8. **Thompson JD, Higgins,DG, Gibson,TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22:4673-80.**

9. **Lipman DJ, Altschul,SF, Kececioglu,JD. A tool for multiple sequence alignment. Proc Natl Acad Sci U S A 1989; 86:4412-5.**

10. **Stoye J, Moulton,V, Dress,AW. DCA: an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. Comput Appl Biosci 1997; 13:625-6.**

11. **Notredame C, Higgins,DG, Heringa,J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000; 302:205-17.**

12. **Gonnet GH, Cohen,MA, Benner,SA. Exhaustive matching of the entire protein sequence database. Science 1992; 256:1443-5.**

13. **Lockless SW, Ranganathan,R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999; 286:295-9.**

14. **Grishin NV. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 1995; 41:675-9.**

15. **Agarwal P, States,DJ. A Bayesian evolutionary distance for parametrically aligned sequences. J Comput Biol 1996; 3:1-17.**

16. **Saitou N, Nei,M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987; 4:406-25.**

17. Studier JA, Keppler,KJ. A note on the neighbor-joining algorithm of Saitou and Nei. Mol Biol Evol 1988; 5:729-31.